Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150005-7

[redacted]

MACHINE INDEXING PROJECT

Interim Report

## SECTION I Summary

The principle mission of the Central Intelligence Agency is the correlation of information within very broad fields. Large numbers of records must be organized so that intelligence specialists can request and receive documents pertaining to a wide range of subjects and to many varying situations, most of which cannot be predicted with certainty at the time that the documents are first received and processed.

The library of documents needed by the Central Intelligence Agency should function as the organization's memory unit. To this end, documents must do more than merely record information; they must also be subject to quick recall to provide information to meet the ever changing requirements for intelligence. To achieve top effectiveness, it must be possible to base the recall of information on new associations of ideas set up to reflect specific situations as they may arise. (For discussion of this problem as seen by the Russians, see Appendix I.)

The first steps to fulfilling this memory function were taken in March, 1947 when punched card methods utilizing conventional IBM equipment together with especially designed and constructed Intellofax machines were successfully applied to the problem. With the increase in volume of documents received and the emergence of unique requirements in certain areas, the need for more versatile and flexible methods and for machines better adapted or especially designed for information searching became apparent. Such machines have been designed and prototypes produced.

The first scanning machine was developed by IBM. An experimental model was exhibited in September 1951 at the Diamond Jubilee Meeting of the American Chemical Society[1].

[1] J. W. Perry and R. S. Casey, "Mechanized Searching", in Encyclopedia of Chemical Technology, Volume 8, Interscience, New York, 1952. (See Appendix XII.)

[redacted]

Application of the new equipment was the subject of a symposium of MIT in June 1952. (See Appendix II.)

Some time before the first announcement to the public, the Central Intelligence Agency was aware of IBM's pioneering development. Late in 1950, a research project was initiated at MIT to explore the usefulness of the new IBM scanning machine for intelligence operations. In this investigation, the indexing was conducted in such a way that all entries pertaining to a given document were punched one after another in a single card. This made it possible to demonstrate, in July 1951, that the ability of the scanning machine to search to combinations of concepts provides a basis for selecting information as needed for varying intelligence requirements. A larger scale project was initiated to develop the possibilities of this new tool.

The new machines search out needed information by a scanning operation comparable to a blind man reading Braille. For this form of machine searching, letters, numerals, and other symbols required to record index entries, are represented by patterns, which may consist of holes in cards, magnetized spots on tape, transparent spots in opaque photographic film or other distinctive marks on some recording medium to be scanned by the machine. Since the number of possible patterns far exceeds the total number of scientific and technical terms and words in general, the new machines impose no restrictions on the range of vocabulary that may be used for indexing. The only limitation is the practical one that the number of index entries assigned to any one document should not be allowed to increase beyond the point of diminishing returns.

The new machines are designed with a number of pattern matching units. Searching and selecting operations may be directed to any one index entry or to combinations of entries. Furthermore, various relations between entries may be specified when setting up the scanning machine prior to initiating a search. This means that a machine search may be tailored with a high degree of precision to meet the needs of intelligence analysts. This dynamic ability to align searches to meet new intelligence assignments is to be contrasted with the inability of a rigid classification system to react to meet changing requirements.

It would have been possible to use the new scanning machine with a conventional punched-card subject code designed for standard punched-card equipment. However, such codes are necessarily constructed on a rigid structure of main and

subclass headings, which do not permit the exploitation of the full potential of the new scanning machine. Therefore, since the scanning machine permits the construction of a more flexible coding system it was decided that such a system should be developed.

Preliminary studies indicated the importance of keeping the indexing system as simple as possible. In working toward this end, careful attention has been given to developing the simplest possible methods of encoding index entries and recording them as patterns searchable by machine. It became evident that certain alterations in the design of the experimental IBM scanning machines would be highly desirable. Conferences with IBM engineers have resulted in plans for production models of the scanning equipment in which the desired changes will be incorporated. (See Appendix III.)

With the new scanning machines, a multiplicity of index entries can be set up for each document and various types of relationships between specific entries encoded and rendered searchable by machine. Thus the code for a specific index entry, such as "bomb", is constructed so as to contain codes for certain selected, related generic terms such as "weapon", "military", "explosive". By using such generic terms as semantic factors in constructing codes, reference points of great effectiveness are provided for defining and conducting searches by machine.

Adequate discriminating power is most important in a machine searching system. For this reason, provision can easily be made to encode characteristic attributes and functional designators. This would make it possible for machine searching to distinguish, for example, between documents in which the same material is used for differnt purposes, e.g., fuel, solvent, food, etc. Something similar to this has already been done, on a limited scale when using certain general action codes, e.g., "Deposits and Resources", "Stockpiles and Reserves, Storage", "Imports-Exports", "Procurement" with other codes for commodities and groups of commodities in the Intelligence Subject Classification.

The program for developing a more flexible indexing system has been directed to meet two basic requirements:

1. The system must be sufficiently simple so that excessive demands or burdens are not imposed on persons who analyze, index, and encode documents.

2. The system must be designed so that machine operations are employed efficiently to accomplish searches as required.

By its very nature, an indexing system must be built up from terminology, that is, from words and from phrases having special meanings. To meet the needs of intelligence the indexing system must make available a wide range of terminology as reference points for defining and conducting searches. Required terminology will be, in part, generic in nature; e.g., "weapon" and, in part, specific, e.g., "rifle", "bomb", "torpedo", "bazooka". The work has been directed (1) to developing methods for organizing terminology into an effective indexing system and (2) to collecting and processing terminology on a mass production basis.

As the first step to developing the new indexing system, terminology from two sources was analyzed and processed along the lines indicated above. These two sources were (1) the Intelligence Subject Classification and (2) outlines of interest prepared by the various divisions of OSI and ORR. In all, about 6500 terms were semantically factored and assigned codes for machine searching.

To provide material for testing the provisional indexing system some 200 documents have been scrutinized by analysts in each of the divisions of OSI and ORR and by information specialists in OCD. In order to indicate what informational features were of interest to them, intelligence experts were requested to underline appropriate words and phrases in the documents or to supply marginal notations. The index entries thus established will next be encoded and the cards punched to permit tests of machine searching.

By July 1, 1953, the provisional system should be sufficiently well developed so that it could be applied to indexing documents for OSI and ORR on a "pilot plant" basis. Such operations could be started as soon as a new scanning machine of appropriate design is available.

During the first period of actual operation, extension of the code to include additional terminology will probably prove necessary. In anticipation of this requirement, some 30,000 terms frequently used for indexing and classifying scientific and technical information have been collected and subjected to preliminary analysis and categorization. Extensive card files have been prepared to facilitate incorporation of these terms into the provisional system as experience during the "pilot plant" test reveals the need for building additional discriminating power into the system.

## RECOMMENDATIONS

1. The development of codes and systems has progressed to the point where encoding and processing of a considerable number of documents is desirable. Some 1000-5000 documents should be incorporated into an expandable "pilot plant" searching system to test the code and to provide a basis for initiating operational application.

2. The pilot system should serve as an initial basis to train operating personnel and to educate potential users as to how to obtain maximum benefit from the system and permit their requirements to guide the establishment of an operational system. Operating procedures, instructions and manuals should be developed during the pilot plant stage.

3. The processing (in particular, the semantic factoring) of collected terminology should proceed at an accelerated rate to provide codes for terms as required for the building and expansion of the "pilot plant" system.



Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150005-7

SECTION II Introduction

## The Nature of Intelligence

"Intelligence," it has been observed, "as an activity is the pursuit of a certain kind of knowledge; as a phenomena, it is the resultant knowledge". In a small way, every person conducts intelligence operations in the course of day-to-day living. Almost any decision on a course of action requires that information be considered and brought to bear on the matter at hand. In everyday life the needed information may be stored in a person's memory as, for example, the telephone number of a friend. At times a very broad range of information may be stored in the human memory and applied to a situation as, for example, in the case of a doctor diagnosing one of the more common ailments. Even for very simple operations, however, it is often necessary to refer to recorded information. An example is the use of the classified section of the telephone directory to find a person or firm to provide a needed service.

In planning the experimental phases of scientific research, a broad range of recorded information--publications, notebooks and reports--usually must be understood and correlated before deciding on a course of action in the laboratory. In such a case the collection of the documents of interest is essentially a clerical task, once the field of research has been reasonably well-defined. The correlation of information preparatory to planning an experimental research program is, in contrast, an intellectual task requiring the highest degree of professional ability.

The use of information in planning and conducting experimental programs in the laboratory finds a close parallel in the use of information by an intelligence specialist to attain insight into obscure situations. In the first place, a broad range of information must be brought to bear if maximum insight is to be attained. Second, the evaluation and correlation of recorded facts requires human skill of the highest order. Finally the recall of needed facts from storage can be made -- and, therefore, should be made -- an essentially clerical task whose accomplishment can be expedited by use of automatic equipment based on modern electronic techniques. The library of records together with the clerical selection and recall system should constitute an external memory able to supply needed information on demand.

## The Nature of Information

The starting point for developing our new machine searching system was recognition of the fact that human understanding of observations, experiments or events involves, first of all, their analysis. For example, various persons have observed recently, circular shining objects whose changing appearance indicated rapid motion at rather high altitudes. This analysis of the phenomena is, however, only the first step to satisfactory "understanding". The next step must be correlation of the new phenomena with previous observations similarly analyzed as to characteristic features. If our analysis is pursued to the theoretical, conceptual level, we often speak of a scientific explanation of the observed phenomena. If correlation of new phenomena proves difficult, a descriptive name such as "flying saucers" may come into general use. The urge to achieve a synthesis with other concepts may go so far as to relate "flying saucers" with space ships of extraterrestrial origin. Correcting an erroneous correlation requires that data adequate to support a correct conclusion be related to the phenomenon under consideration.

The analysis of individual events and their correlation in terms of common features has characterized the development of science and the scientific method. The same two basic steps are also involved in converting uncoordinated information into finished intelligence. Or, to state a general conclusion, the analysis of information is the first step in its conversion into knowledge.

## The Role of Machine Searching

An essential preliminary operation in the correlation of recorded facts is the selection of those items of information having features in common. The establishment of a list of entities, processes, circumstances and results--in short, a list of features--that can serve as reference points for defining searching operations, is the first step to advantageous use of certain mechanical and electronic devices to select information as needed. During recent years, the possibilities of using punched cards, in particular hand-sorted punched cards, have been investigated by many persons. (See Appendix XII.) Automatic punched-card machines designed for accounting and business applications have been found to have limitations that restrict their usefulness for searching and correlating non-numerical data for producing intelligence. (See Appendix XIII.)

The scanning machines with which this report is concerned were developed to overcome the limitations of previously available equipment. The speed of operation of the new scanning equipment makes it possible to search all the entries in a lengthy index. This means that alphabetizing index entries in the usual way is not only superfluous but disadvantageous. Machine searching is, in fact, rendered more effective by holding the entries pertaining to a given document together as a block. This facilitates selection when the search is defined as involving a combination of entries.

The flexibility of this form of machine searching can be illustrated by a simple example. A search can be directed readily to any one index entry, such as "barley", "wheat", "rust", "rainfall", "Siberia". Alternately, selection may be based on a combination such as "wheat", "rust", and "Siberia", or "wheat", "rainfall" and "Siberia" or "barley". Other types of combinations available for use if needed will be discussed in connection with review of the operating characteristics of the machine.

SECTION III    New Electronic Equipment

## Types of Patterns for Searching

Recent advances in electronic techniques make it possible to construct a variety of scanning machines each of which searches and selects by matching patterns at high rates of speed. Depending on which type of scanning machine is chosen, the patterns consist of any one of several types of distinctive marks impressed on a suitable recording medium. With IBM cards, the patterns are built up of punched holes. Spots of various types provide other possibilities. Thus patterns for machine searching may consist of magnetized spots on tape, transparent spots in an opaque photographic film, or lack spots printed on cards. Any type of pattern is equally effective for recording letters, numbers or other symbols employed to spell out index entries.

## Pattern Matching

In order to search for combinations of index entries, it is, of course, essential that the searching machine be able to detect several patterns within a block of entries pertaining to a single document. For simplicity, discussion of pattern matching methods will be in terms of a single meaningful pattern.

The first experimental scanning machine constructed by IBM for information searching detected patterns photo-electrically. Preparatory to conducting a search, a question card was punched with a pattern that was the complement of the one to be sought in the file cards. In conducting the search the question card was held in a stationary position and the cards from the file being searched were moved longitudinally past the question card. The light source and the photo-electric cells were arranged in such a fashion that when a pattern in one of the file cards passed through a matching position with the complementary pattern in the question card, the light was momentarily cut off from one of the photocells. This use of a photocell to detect a matching condition between a pattern and its complement can also be employed to detect a pattern recorded as transparent and opaque spots in a photographic film.

A radically different procedure for detecting a specified pattern involves a two-step process. First, the patterns being scanned are converted into distinctive sets of electrical pulses. When the patterns are recorded as magnetized spots on tape, reading

heads designed for digital computers may be used. A video tube may be used for the same purpose, when the patterns are recorded as black spots printed on cards. With punched cards, it might be possible to use either brush contacts or photo-electric means. The latter is preferred at present by IBM engineers. (This use of photo tubes with IBM cards must not be confused with the matching by the blackout method already described.)

Once the patterns in the recording medium have been converted into sets of photelectric pulses, detection of some one set may be accomplished by an electronic comparator unit. Before starting the search the machine operator must set this unit so that it will respond only to the set of pulses corresponding to the desired pattern and consequently to the index entry to which the search is directed.

The two-step method of detecting patterns offers important advantages over the blackout method originally employed in the first experimental scanning machine constructed by IBM. Conversion of recorded patterns into sets of pulses can be accomplished with high reliability by various reading devices already mentioned above. Furthermore, comparator circuits arranged in parallel have decisive advantages over a bank of photocells operating on the blackout principle. Most important is the possibility of achieving extensive simplification in both the indexing of information and in the codes used to represent index entries. Such simplification can be achieved independently of the recording media (punched cards, photographic film, magnetic tape, etc).

With searching systems employing IBM cards, the two-step method of detecting patterns offers certain specific advantages. One of these is the possibility of conveniently using standard punching with IBM cards. As discussed in detail in Appendix III, the production models of the IBM scanner will be designed to work with standard punching. This opens the door to using the new scanner with IBM files already in existence.

The two-step process for detecting patterns can be employed with tabulating or other cards in which punching has been replaced by black spots printed on the cards. Patterns of spots are converted into sets of pulses by a video tube. In this system--described in more detail in Appendix IV--one side of the card can be used for an abstract, bibliographical data or the like.

## The Selecting Operation

Regardless of the medium in which a given pattern is recorded, and regardless also of the corresponding means used for detecting patterns, the identification of a designated pattern can be made to generate an electrical pulse which can then be used to activate a selecting device. When working with files of cards, the selecting operation may be the physical removal of the card containing the desired pattern from the file. Once a group of cards has been selected out, the scanning machine may be reset and the selected group submitted to a further selecting operation. This possibility of conducting successive selecting operations in series without being compelled to scan the entire file may prove an important factor in favor of using cards or other types of records consisting of discreet units rather than magnetic tape or other continuous media. With the latter the selecting operation which corresponds to removal of individual cards from a file might consist of reading off the serial numbers of the documents identified and recording these numbers for example, on a separate output tape.

## Searching to Combinations of Entries

Modern electronic techniques permit the searching operation to be defined in a more complex fashion than corresponds to matching a single pattern. This purpose is best served by a multiplicity of pattern detecting units, preferably comparator circuits. Each detecting unit is set so as to respond to some one pattern, which in turn corresponds to one of the index entries used to define the scope of a search being conducted. As already noted, a matching condition generates an electrical pulse in the output terminals of one of the detecting units. Establishing appropriate auxiliary relay circuits between these output terminals makes it possible to condition the machine so that a selecting operation is performed only when a prespecified relationship exists between the detected patterns (i.e., between the corresponding index entries). With the aid of an appropriately designed plugboard, it is a simple matter to condition the scanning so that selection occurs only if each and every one of a set of patterns has been detected. Using letters to designate the patterns (or their corresponding index entries) such a search may be symbolically represented as follows and is termed a logical product--

$A \cdot B \cdot C \cdot D$ (in the case of four patterns)

It is equally easy to condition the scanning so that selection is based on the presence of any one of several factors and such



Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150005-7

selection is said to correspond to the logical sum, symbolized, for example, as:

$$A + B + C + D$$

A further simple possibility is to base selection on the presence of one pattern and the absence of another. This type of selection is termed a logical difference and symbolized as:

$$A - B$$

Finally, more complicated relationships can be specified in defining the scope of the search. One might specify, for example, that selection is to occur if any one of a group of several patterns be present together with any one of another group. This might be represented symbolically by:

$$(A + B + C) \cdot (D + E + F)$$

Even more complicated relationships between patterns may be specified, such as:

$$(A + B) \cdot (C - D) + (E - F)$$

The possibility of conducting searches in the manner exemplified by these equations has not been available to intelligence operations in the past.

## Rates of Scanning and Selecting Operations

Regardless of whether punched cards, magnetic tape or some other medium may be used for recording meaningful patterns the electronic recognition of a desired predetermined pattern can be accomplished at an extremely high rate of speed. With the new experimental IBM machines, it is possible to scan 50,000 patterns per minute. It would be possible, using electronic techniques already developed, to construct a tape machine which could scan at least 2,000,000 patterns per minute. These searching rates, already attained or attainable, are such that the entire set of index entries used to indicate important aspects of the subject matter of a file of documents can be scanned in a reasonable time. The scanning operation itself is comparable to a blind man using his fingertips to detect the raised patterns of dots which constitute representation of words in Braille. The result of the scanning operation is comparable to having a highly reliable, though unimaginative, clerk inspect all the entries in a comprehensive subject index at a very high rate of speed.

## SECTION IV Establishing Teamwork Between Men and Machines

### Capabilities and Limitations of Machines

As the previous discussion has pointed out, electronic machines which employ pattern matching for identifying and selecting purposes are able to perform automatically clerical searching operations of considerable complexity. In considering how best to use such equipment in an intelligence organization, we must keep in mind certain basic limitations as to what machines can accomplish. Their ability is outstanding in performing well-defined routine operations without fatigue, and at a high rate of speed. Aside from the searching of indexed information, the new scanning equipment could be used advantageously to select and correlate information from various files already established or files based on titles of published papers (see Appendix V ). The specifications of the scanning equipment are provided in Appendix III so that interested offices may determine how the new machines may be used to advantage.

Machines are devoid of any power to evaluate or interpret. This means that the usefulness of searching machines is limited to identifying what documents are of pertinent interest to a given question. The interpretation of the significance of the subject matter of documents and the correlation of information contained in them requires the vastly superior ability of human intelligence. The identifying and selecting operations performed by machines can, however, expedite the work of intelligence analysts very effectively by making it unnecessary for them to review large masses of material not pertinent to the problem at hand. Machine searching enables analysts to devote their time and effort more fully to the interpretation and correlation of pertinent information rather than to inspection and rejection of information of no interest to a given problem. Our goal is to establish effective teamwork between machine performance of routine clerical tasks and human experts interpreting selected pertinent information and creating intelligence.

### Discriminating Power of Machine Systems

The basic problem in designing a machine indexing system for intelligence purposes is to make it as simple as possible and at the same time provide adequate discriminating power to select

documents for the analyst's consideration in preparing his reports. More specifically the immediate problem is to develop indexing techniques that will exploit most efficiently the potentialities of the new searching machines for intelligence requirements. It is, of course, necessary to avoid setting up an overly complex indexing system which--though it might be very efficient when viewed exclusively from the viewpoint of machine operations--would nevertheless make excessive demands of persons charged with the tasks of analyzing documents and encoding index entries.

In developing the new indexing system, there are two courses of action that might be followed. One is to construct an excessively complex system, applying it experimentally, and then simplify it to fit actual needs. This is an approach likely to prove attractive to theoretically-minded persons who have not had practical experience in the field of indexing and coding. We have preferred a different approach, which involves first setting up a simple system for testing and operation during which experimental results would provide the basis for building into the system such additional discriminating power as might be necessary to meet the needs of intelligence specialists.

## Indexing for Machine Searching

A basic consideration in designing the new indexing system is economy of effort on the part of persons who inspect documents to determine and to indicate what aspects are of importance for intelligence purposes. These essential tasks can be best accomplished, during the immediate future at least, by having an expert in the subject field underline appropriate words and phrases in the document and also provide supplementary marginal notations when necessary. Eventually, a suitably designed machine may be able to perform this task sufficiently well to serve certain purposes, such as document screening. Until that time, however, this first step will require considerable human effort. To minimize this, it is anticipated that marking documents for indexing may be made a routine part of the analysts' reading of incoming documents.

## Encoding Terms for Machine Searching

Once words and phrases have been selected as index entries, the next step is to encode them and record them, e.g., by punching cards, so that machine searching operations may be accomplished. The simplest possible form of encoding would involve nothing more than using appropriately selected patterns of holes, or such,

as means for recording letters and numerals. In this approach, the recording patterns would spell out the index entries in exactly the same way that embossed patterns of dots are used to spell out words in Braille. This simple form of encoding can be accomplished automatically by machines already in existence. Depressing a key corresponding to a given letter or numeral causes the corresponding pattern to be punched in a card or otherwise appropriately recorded. In other words, an operation very similar to typing suffices to render the index entries immediately searchable by machine. Once the index entries have been recorded as appropriate patterns, searching operations may be directed to any word or combination of words. This approach has the unquestionable advantage of simplicity. On the other hand, it suffers from certain disadvantages. One of these, which is closely related to problems encountered in using conventional subject indexes, is the necessity for the operator of the machine to set it so that searching and selecting is based on the right words. Synonyms form one obvious source of difficulty in this connection. Since the machine operates by pattern matching, a search directed to a pattern representing "mercury" will not produce a positive response when the pattern for "quicksilver" is encountered. Near-synonyms and terminology having overlapping areas of meaning would result in similar and perhaps even more vexing limitations on the usefulness of this simple approach, even though it can almost certainly provide considerable aid in rapid screening of documents. (For tests with this simplest form of coding, see Appendix V.)

The above mentioned, simplest form of encoding is also limited as to usefulness by another factor. This limitation becomes apparent if we consider the situation that would result if the patterns were used to spell out the names of insects, e.g., "mosquito", and then a search is required for all insects indexed as causing damage to wheat in the Ukraine. Before the machine could be set to conduct this search a list of all the insects that might be involved in damaging wheat in the Ukraine would have to be compiled. A theoretical alternative--completely impractical in terms of machine operations--would be to base the search on a list of all known insects. It would also not be satisfactory to attempt to meet the search requirements by selecting documents on damage to wheat in the Ukraine as this would result in simultaneous selection of all documents on damage due to plant diseases, unfavorable weather conditions, weeds, and the like. Along with documents of pertinent interest, the analyst would almost certainly have to cope with an excessive amount of extraneous material. It may, therefore, be concluded that indexing for machine searching can achieve a very large measure of added effectiveness if generic terms, such as "insect", are provided as entries to accompany specific terms, such as "mosquito".

It is perhaps obvious that success in indexing for machine searching will depend in large degree on how terminology is employed to provide reference points for selecting operations to be conducted by machine. It might be possible, theoretically at least, to require the indexer always to use only one of two or more synonyms or near-synonyms. This would probably prove much more practical than requiring that the indexer provide generic terms, such as "insect", every time a specific term, such as "mosquito", is required as an index entry. Remembering--and applying--appropriate generic terms for a large number of specific terms would be excessively burdensome. The rate of indexing would be slowed to an intolerable degree.

It is perhaps clear that machine searching can achieve a high level of effectiveness only when a solution is found for the double-barreled problem of relieving the indexer of excessive burdens and of simultaneously using generic terminology as effective means for rendering information searchable by machine operations. The key to this problem is the systematization of terminology on the basis of semantic relationships as between "insect" and "mosquito", for example. The considerations underlying the establishment of such semantic relationships between terms are reviewed in Appendix VI. The method whereby the generic terms are rendered effective is to set them up as component parts of the codes of specific terms. The generic terms related in this way to a given specific term are referred to as the latter's semantic factors. When a specific term is used as an index entry, encoding the latter makes its semantic factors available as generic terms to serve as reference points for machine searching.

There were other less effective approaches to the problem of employing generic terms to define searches. It would have been possible to use conventional classification with the new scanning equipment. Since such conventional classification systems are set up to operate with a fixed array of pigeon holes, they are characterized by a high degree of ridigity. The same is true of the Intelligence Subject Classification which was designed to work with standard IBM tabulating equipment. Rigidity in the system used for analyzing information would have meant that the inherent flexibility of the new machines would have been exploited only to a very limited degree. If the new developments in equipment were to be used effectively there had to be a corresponding advance in indexing methodology. In order to avoid burdening indexers unduly, the requisite advance in methodology required that much care and thought be devoted to code construction.

Processing and systematizing of terminology and constructing of codes have been carried out on a mass production basis. First,

subject heading lists, classification systems and indexes were collected as sources of terminology (See Appendix VIII). Next a file of some 30,000 "Keysort" cards was prepared. Individual terms were entered on each card in this file, together with dictionary definitions of the various terms. Supplementary notes concerning recent trends in usage of terms were included particularly when concerned with rapidly developing fields such as electronics. The cards were then punched to indicate categorization of the terms both as to general type (e.g., Processes) and field of use (e.g., Chemistry). For convenience in semantic factoring, more specific grouping of the terms as to area of use (e.g., Food and Fermentation) was carried out for source terms. An IBM file was also prepared for automatic manipulation of terms and for rapid preparation of listings. (For a description of these files, their preparation, and use in semantic factoring and code construction, see Appendix IX.)

Some 6500 terms found in the ISC and supplied by analysts as outlines of interest have been encoded ready for use in the indexing of documents for machine searching. To illustrate how the new indexing system operates, the analysis and encoding of a typical document will now be described.

## Document Analysis and Encoding

Intelligence documents coming into CIA are perused in order to decide the routing of each document to the various offices and divisions of the Agency. The analysts of OSI receive documents of probable interest. The analysts accomplish the first step in the indexing procedure by underlining words and phrases that indicate subjects of interest. If necessary, additional subjects are indicated by marginal annotations. In experimental testing of the new indexing system, representatives of each of the divisions of ORR and OSI check the document to ensure that all points of view of importance are designated by appropriate words and phrases. Experience in processing 200 documents indicate that this supplementary checking results in relatively few additional index entries being set up.

The encoding of index entries procedes as follows:--

A. Work Sheet: A work sheet is prepared for each document (See Figure A), and the following information entered:

1. Document number--This is a new serial number assigned to each document on arrival.

There is no relation to the serial number assigned by OCD, since documents other than those processed by the Library may be included.

2. OCD number.--If the document has been processed by OCD, the identification number is entered here. If it is a CIA report, that report number is recorded.

3. Date of information: month, day, and year, if given.

4. Date of report: month, day, and year, if given.

5. Security Classification.

6. Terms, phrases, and subjects as marked and annoted by analysts. Included here are:

a. Any geographical locations of interest, whether city, country, or any other subdivision.

b. Any personality names of interest.

c. Any names of institutes or companies of interest.

d. Any index entries referring to subject matter as indicated by the underlined terms or phrases and by marginal annotations.

7. Under the column "Type" is entered the type of information represented by a given term or annotation (for example, G = geographical location; I = institute, company, etc.; S = other subject matter).

8. The columns under "Division Marking" identify the particular group or groups in CIA which found a particular term of interest. This information is of value from the point of view of future research and development based on the statistical data generated.

9. Codes are entered in the last column of the work sheet. The encoder refers to the

[redacted]

FIGURE A. Experimental Code Sheet — Security Information

MACHINE INDEXING PROJECT
Code Sheet

| | | Doc. # M - 32 | OCD. - - - | TYPE | Division Marking C | P | N | W | M | V | B | O | L | Other | CODE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Doc. # M - 32 | OCD. - - - | | | | | | | | | | | | |
| 2 | Date (a) - - -<br>(b) 110152 | | | | | | | | | | | | | |
| 3 | Source: United Nations Bulletin | | | | | | | | | | | | | SUNB |
| 4 | Security: Unclassified | | | | | | | | | | | | | 9 |
| 5 | | | | | | | | | | | | | | |
| 6 | Marked Terms<br>World | | G | | | | | - | | | | | | LA |
| | W. H. O. | | I | | | | | - | | | | | | I:WHO |
| | World Health Assembly | | I | | | | | - | | | | | | I:WHA |
| | Quarantine | | S | | | | | - | | | | | . | :SICE:06:<br>CONU |
| | Station | | S | | | | | - | | | | | . | :LOCA:06 |

Terms listed by:
Approved:
Key punch:

[redacted]

Security Information

[redacted]

Security Information

English-to-code dictionary to find the code equivalent to the indexed term and enters it on the work sheet. (See Appendix XI for description of the code dictionary. The code dictionaries are available in Room 1348 M. Please see [redacted]

25X1A9a

25X1A9a

B. Key Punching: A tabulating card (or cards) is prepared for each document processed. The document numbers and codes for the dates of information and report, source and security classification are punched in "fixed fields". The remaining codes are punched one following the other as they appear on the experimental code sheet.

## Searching of Encoded Documents

Previous discussion has described how documents may be analyzed, index entries set up and the latter encoded in such a way as to make semantic factors available for defining searches. Each factor--which, as noted already, is a generic concept--is represented by a four letter code. Any one of these four letter codes may be used as a reference point for defining the scope of a search. Example; a search directed to MACI (code for the semantic factor "machine, device, apparatus") will result in a positive response on the part of the scanning machine whenever MACI appears as a factor in a specific term used to index a document. If the search is directed to codes containing both MACI and MESU (the code for "measure") then there will be a positive response whenever a specific term designating a measuring device or apparatus has been used to index a document. Such measuring devices would include, for example: densitometer, clock, buret, potentiometer, watch.

Specification of an additional factor such as "time" (encoded in TIME) would result in positive response only when a document was indexed by using a term such as "clock" or "watch", designating a time measuring device (corresponding codes MACI, MESU, TIME). In this way inclusion of semantic factors in the codes for specific index entries permit generic terms to be used in conducting searches. It is important to note that two or more semantic factors may be used to establish a single broad class of items (e.g., time measuring devices) which then becomes one dimension in defining the scope of search.

Several such dimensions may be set up simultaneously. For example, in addition to time measuring devices as one dimension of search, we might establish explosive weapons, encoded by using the semantic factors "explosive" (POLI) and weapon (WEPO) as another dimension. If this were done, then the search would result



[redacted]

Security Information

in a positive response for documents in which one search dimension is provided by some time measuring device (e.g., clock, watch) and another dimension by an explosive weapon (e.g., bomb, torpedo).

In considering the possibilities inherent in the new indexing and searching system, it must be recalled that various relationships may be set up between search dimensions. If these are denoted by letter, we have three basic possibilities:

(1) A · B (there must be a positive response to both A and B, so-called logical product).

(2) A / B (there must be a positive response either to A or to B, or to both, so-called logical sum).

(3) A - B (there must be a positive response to A but not to B, so-called logical difference).

From these, more complex configurations may be built up from a series of dimensions used to define a search.

Examples of such possibilities are:

(A · B) / (C - D) · E · F

(A - B - C) / (C · D - E · F)

Skill in exploiting these possibilities can probably best be developed by operational experience.

## Role Indicators and Further Refinement of Machine Searching

Discussion up to this point has emphasized the use of appropriately selected terminology as index entries which, once encoded and recorded, can be searched by machines. It has been pointed out that machine searching can be rendered more effective and efficient by building into the code certain terminology relationships, such as whole-part, genus-species or others involved in semantic factoring. By taking advantage of these possibilities it is possible to design a simple code, whose use involves a minimum of human effort and which enables machine searching operations to achieve a high degree of discrimination. An additional measure of discrimination becomes possible by using

appropriate symbols in indicating the role of a given entity, process, attribute, or the like. Thus, to take a simple example, by attaching appropriate role indicators to "man" and to "dog", it would be possible to distinguish between "man bites dog" and "dog bites man". Similarly, special symbols might be used to indicate the direction of movement of goods, persons, army units, etc. from one place to another. Another possible use for special symbols attachable to any subject entry such as "typewriter" is to indicate whether the item in question is being manufactured, used, tested, developed, bought, sold, etc. The OCD system's successful use of slash marks has already demonstrated the usefulness of this device.

For a more detailed discussion of role indicators, see Appendix X.

## CONCLUSION

A new system for indexing information and for searching large files has been developed. The system is based on specially designed machines whose operating characteristics place no limitation on the range of terminology available for indexing purposes. Searches may be directed with equal speed and facility to any one index term or to their combinations. The flexibility of the new system was developed so that it would be able to meet the unusually demanding requirements of OSI and ORR. To ensure economy of operation the new indexing system has been designed so as to require a minimum of human effort, particularly on the part of experts who analyze the subject matter of documents.